

A Family of Blockwise One-Factor Distributions for Modelling High-Dimensional Binary Data

Matthieu Marbac and Mohammed Sedki

November 5, 2015

Abstract

We introduce a new family of one factor distributions for high-dimensional binary data. The model provides an explicit probability for each event, thus avoiding the numeric approximations often made by existing methods. Model interpretation is easy since each variable is described by two continuous parameters (corresponding to its marginal probability and to its strength of dependency with the other variables) and by one binary parameter (defining if the dependencies are positive or negative). An extension of this new model is proposed by assuming that the variables are split into independent blocks which follow the new one factor distribution. Parameter estimation is performed by the inference margin procedure where the second step is achieved by an expectation-maximization algorithm. Model selection is carried out by a deterministic approach which strongly reduces the number of competing models. This approach uses a hierarchical ascendant classification of the variables based on the empirical version of Cramer's V for selecting a narrow subset of models. The consistency of such procedure is shown. The new model is evaluated on numerical experiments and on a real data set. The procedure is implemented in the R package *MvBinary* available on CRAN.

Keywords: Binary data, EM algorithm, High-dimensional data, IFM procedure, Model selection, One-factor copulas.

1 Introduction

Binary data are increasingly emerging in various research fields, particularly in economics, psychometrics or in life sciences (Cox and Snell, 1989; Collett, 2002). To carry out statistical inference, it is important to have at hand flexible distributions for such data. However, there is a shortage of multivariate distributions for binary data (Genest and Nešlehová, 2007). Indeed, many approaches have been developed by considering that the binary variables are responses of several explanatory variables (Glonek and McCullagh, 1995; Nikoloulopoulos and Karlis, 2008; Genest et al., 2013). However, these models cannot manage data composed with only binary variables.

Since binary variables are easily accessible and poorly discriminative, the binary data sets are often composed of many variables. Thus, the modelling of high-dimensional binary data is an important issue. Moreover, classical models suffer from the *curse*

of *dimensionality* since they involve too many parameters (Bellman, 1957). Therefore, specific distributions should be introduced to manage such data.

Many authors have been interested in defining the properties of a multivariate distribution which permit an easy interpretation and inference (Nikoloulopoulos and Karlis, 2009; Panagiotelis et al., 2012). Thus, Nikoloulopoulos (2013) lists the five following features that define a distribution with good properties: (F1) Wide range of dependence, allowing both positive and negative dependence; (F2) Flexible dependence, meaning that the number of bivariate marginals is (approximately) equal to the number of dependence parameters; (F3) Computationally feasible cumulative distribution function for likelihood estimation; (F4) Closure property under marginalization, meaning that lower-order marginals belong to the same parametric family; (F5) No joint constraints for the dependence parameters, meaning that the use of covariate functions for the dependence parameters is straightforward.

The modelling by *dependency trees* (Chow and Liu, 1968) is a pioneer approach for assessing the distribution of binary variables. A strength of this method is the easy maximization of likelihood function by the Kruskal algorithm (Kruskal, 1956), which estimates the tree of minimal length. Although the tree structure leads to benefits (estimation, visualisation and interpretation), it is limited to simple dependency relations. Moreover, it does not provide parameters for measuring the strength of the dependencies between two variables.

A naïve approach for modelling binary variables is to use a product of Bernoulli distributions. However, in spite of the parsimony induced by the independence assumption, this approach leads to severe biases when variables are dependent. Thus, a *mixture model* with conditional independence assumption can capture the main dependencies (Goodman, 1974). Celeux and Govaert (1991) propose different parsimonious models to deal with high-dimensional data. However, this mixture-based method suffers primarily from a lack of interpretation of dependencies. Indeed, there is no parameters for directly reflecting the strength of the dependencies between variables.

The *quadratic exponential binary distribution* (Cox, 1972) is considered as the binary version of the multivariate Gaussian distribution. However, this model does not retain its exact form under marginalization, but closure under marginalization can be achieved approximately (Cox and Wermuth, 1994). This model is not really suitable for high-dimensional data since it implies a quadratic number of parameters.

The modelling of spatial binary data can be achieved by latent Gaussian Markov Random Fields (Pettitt et al., 2002; Weir and Pettitt, 2000) or lattice-based Markov Random Fields like the Ising model (Gaetan et al., 2010). These approaches can deal with high-dimensional data since they have the Markov properties. However, their using for non-spatial data is not really doable except when the model at hand is known. Indeed, this approach requires to define the neighbourhood of each site. The combinatorial issue of model selection also prevents the use of such approaches when data are non-spatial.

The general approach to model multivariate distributions is to use copulas. Indeed, *copulas* (Nelsen, 2006; Joe, 1997) can be used to build a multivariate model by defining, on the one hand, the one-dimensional marginal distributions, and, on the other, the dependency structure. Among the copulas, the *Gaussian* and the *Student* ones are very popular since they model the pairwise dependencies. However, their likelihood has not a closed form when the variables are discrete. It can be approached by numerical integrations

which is not doable for high-dimensional data. Moreover, they require a quadratic number of parameters which leads to the curse of dimensionality for high-dimensional data. Alternatively the *Archimedean copulas* are relevant to reduce the number of parameters since they use a single parameter to model the dependencies between all the variables. Thus, this parameter characterizes a general dependency over the whole variables but it also limits the interpretation. For instance, positive and negative dependencies cannot be modelled simultaneously. Moreover, the evaluation of the likelihood requires the evaluation of an exponential number of terms, so it is not doable for high-dimensional data. Finally, *vine copulas* (Kurowicka, 2011) are a powerful alternative since they allow the specification of a joint distribution on d variables with given margins by specifying $\binom{d}{2}$ bivariate copula and conditional copula. Note that the vine copulas generalize and increase the flexibility with respect to the dependencies trees.

The *one factor copulas* (Knott and Bartholomew, 1999) enable to reduce the number of parameters and thus to deal with high-dimensional data. This approach assumes that the dependencies between the observed variables are explained by a continuous latent variable. This approach can be used for modelling continuous data variables (Krupskii and Joe, 2015), extreme-value continuous data (Mazo et al., 2015) or ordinal data (Nikoloulopoulos and Joe, 2013).

In this work, we introduce a new family of one factor distributions that can be written as a specific one factor copula. For modelling more complex dependency structures, we extend this family by allowing a partition of the set of observed variables into independent blocks, where each block follows the new one factor distribution. The resulting family respects the five features listed by Nikoloulopoulos (2013). According to this specific distribution, each variable is described by three parameters: a continuous parameter indicating its marginal probability, a continuous parameter indicating the strength of the dependency with the rest of variables of the block (through the latent variable) and a discrete parameter indicating if the dependency is positive or negative.

Since the proposed distribution is a specific copula for discrete data, parameter inference is achieved by a two step procedure named Inference Function for Margin (IFM, see Joe (1997, 2005)). Model selection consists in finding the best partition of the variables into blocks according to the Bayesian Information Criterion (BIC; Schwarz et al. (1978); Neath and Cavanaugh (2012)). Although this information criterion is defined with the maximum likelihood estimates, an extension has been proposed with the parameter estimates resulting from the IFM (Gao and Song, 2010). For high-dimensional data, an exhaustive approach computing the BIC for each possible model is not doable. Therefore, we propose a deterministic two step procedure for the model selection. First, a small subset of models is extracted from the whole competing models by a deterministic procedure based on a Hierarchical Ascendant Classification (HAC) of the variables by using their empirical Cramer's V. Second, the BIC is computed for the models belonging to this subset and the model maximizing this criteria is returned. We show that this approach is asymptotically consistent. Indeed, Metropolis-Hastings algorithm (Robert and Casella, 2004) is used for detecting the model maximizing the BIC criterion. Alternatively, a Metropolis-Hastings algorithm (Robert and Casella, 2004) can also be used for detecting the model maximizing the BIC criterion. However, we numerically show that the deterministic procedure obtains similar results, in a strongly reduced computing time, as the stochastic one. Therefore, we advise to use the deterministic procedure.

The paper is organised as follows. Section 2 introduces the new family of the specific one factor distributions per independent blocks. Section 3 presents the parameter inference with the IFM procedure. The model selection issue is detailed in Section 4. Section 5 numerically compares both model selection procedures. Section 6 illustrates the approach on a real data set. Section 7 concludes this work. All the mathematical proofs are in appendix. The R package MvBinary implements the proposed method and contains the real data set. It is available on CRAN and the url <http://mvbinary.r-forge.r-project.org/> proposes a tutorial for reproducing the application described in Section 6.

2 Multivariate distribution of binary variables

2.1 Blocks of independent variables

The aim is to model the distribution of the d -variate binary vector $\mathbf{X} = (X_1, \dots, X_d)$. Variables are grouped into B independent blocks for modelling different kinds of dependencies. Thus, the vector $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d)$ determines the block of each variables since $\omega_j = b$ indicates that X_j is assigned to block b with $1 \leq b \leq B$. Therefore, independence between blocks implies

$$\forall 1 \leq j \leq j' \leq d : \omega_j \neq \omega_{j'} \implies X_j \perp X_{j'}. \quad (1)$$

Vector $\boldsymbol{\omega}$ defines a model which is unknown and which has to be inferred from the data. Variables affiliated to block b are mutually dependent and are denoted by $\mathbf{X}_{\{b\}} = (X_j : \omega_j = b)$. Obviously, this approach allows to model dependencies between all the variables (*i.e.* $B = 1$ then $\omega_j = 1$ for all $1 \leq j \leq d$) or independence between all the variables (*i.e.* $B = d$ then $\omega_j = j$ for all $1 \leq j \leq d$). The probability mass function (pmf) of the realisation $\mathbf{x} = (x_1, \dots, x_d)$ is

$$p(\mathbf{x}|\boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{b=1}^B p(\mathbf{x}_{\{b\}}|\boldsymbol{\theta}_b), \quad (2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_b; b = 1, \dots, B)$ groups the model parameters, where $\boldsymbol{\theta}_b$ groups the parameters of the variables of block b . Finally, $p(\cdot|\boldsymbol{\theta}_b)$ is the pmf of variables affiliated to block b and each block is assumed to follow the one-factor distribution described in the following.

2.2 One-factor distribution per blocks

2.2.1 Conditional block distribution

In block b , dependencies between variables are characterised through a random continuous variable U_b which follows a uniform distribution on $[0, 1]$. More precisely, variables of block b are independent conditionally on U_b . So, the pmf of variables affiliated to block b is

$$p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) = \prod_{j \in \Omega_b} p(x_j|u_b, \boldsymbol{\theta}_j), \quad (3)$$

where $\boldsymbol{\theta}_b = (\boldsymbol{\theta}_j; j \in \Omega_b)$, $\boldsymbol{\theta}_j$ denotes the parameters related to variable X_j detailed below, and where $\Omega_b = \{j : \omega_j = b\}$ is the set of the indices of the variables of block b . Therefore, the specific conditional distribution of $\mathbf{x}_{\{b\}}$ is assumed to be a product of Bernoulli distributions whose parameters are defined according to the value of u_b . Indeed, for $j \in \Omega_b$

$$p(x_j|u_b, \boldsymbol{\theta}_j) = p_j^{x_j} (1 - p_j)^{1-x_j} \text{ with } p_j = (1 - \varepsilon_j)\alpha_j + \varepsilon_j \mathbb{1}_{\{u_b < \alpha_j\}} \mathbb{1}_{\{u_b > 1 - \alpha_j\}}^{1-\delta_j}, \quad (4)$$

where $\boldsymbol{\theta}_j = (\alpha_j, \varepsilon_j, \delta_j)$ groups the parameters related to variable X_j where:

- the continuous parameter $\alpha_j \in (0, 1)$ corresponds to the marginal probability that $X_j = 1$ since one can easily verify that for $j \in \Omega_b$, $\int_0^1 p(X_j = 1|u_b, \boldsymbol{\theta}_j) du_b = \alpha_j$,
- the continuous parameter $\varepsilon_j \in (0, 1)$ reflects the dependency strength between X_j and the other variables of block j since the stronger the ε_j , the more correlated are the variables of the block (see Proposition 2.3),
- the binary parameter $\delta_j \in \{0, 1\}$ indicates the nature of the dependency, since $\delta_j = 1$ if the observed variable is positively dependent with the latent variable and $\delta_j = 0$ otherwise. Thus, two variables X_j and $X_{j'}$ affiliated to the same block (*i.e.* $\omega_j = \omega_{j'}$) are positively correlated if $\delta_j = \delta_{j'}$ and they are negatively correlated if $\delta_j = 1 - \delta_{j'}$.

Note that the model identifiability is discussed in the next section.

The parametrization of (4) is convenient for the model interpretation. However, we introduce the following new parametrization which simplifies the likelihood computation. Conditionally on u_{ω_j} , x_j follows a Bernoulli distribution whose the parameters are only determined by a relation between u_{ω_j} and real $\beta_j = \alpha_j^{\delta_j} (1 - \alpha_j)^{1-\delta_j}$ which corresponds to the marginal probability that $X_j = \delta_j$. Indeed, for $u_{\omega_j} \in [0, \beta_j]$, the conditional distribution $X_j|u_{\omega_j}, \boldsymbol{\theta}_j$ is a Bernoulli distribution $\mathcal{B}(\lambda_j)$ where $\lambda_j = (1 - \varepsilon_j)\alpha_j + \varepsilon_j\delta_j$. Moreover, for $u_{\omega_j} \in [\beta_j, 1]$, the conditional distribution $X_j|u_b, \boldsymbol{\theta}_j$ is a Bernoulli $\mathcal{B}(\nu_j)$ where $\nu_j = (1 - \varepsilon_j)\alpha_j + \varepsilon_j(1 - \delta_j)$. Thus, (4) can be summarized as follows

$$p(x_j|u_b, \boldsymbol{\theta}_j) = \begin{cases} \lambda_j^{x_j} (1 - \lambda_j)^{1-x_j} & \text{if } 0 \leq u_b < \beta_j \\ \nu_j^{x_j} (1 - \nu_j)^{1-x_j} & \text{if } \beta_j \leq u_b < 1 \end{cases} \quad (5)$$

2.2.2 Marginal block distribution

Obviously, the realizations u_b are not observed, but the distribution of the observed variables $\mathbf{X}_{\{b\}}$ results from the marginal distribution of the pair $(\mathbf{X}_{\{b\}}, U_b)$. So, the pmf of $\mathbf{x}_{\{b\}}$ is defined by

$$p(\mathbf{x}_{\{b\}}|\boldsymbol{\theta}_b) = \int_0^1 p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) du_b. \quad (6)$$

We now describe the properties of the block distribution. All proofs are given in Appendix A. For respecting the feature (F3) of Nikolouloupoulos (2013) and for dealing with high-dimensional data, the block distribution needs to have a closed form. This explicit pmf is detailed in the following proposition.

Proposition 2.1 (Explicit distribution) Let σ_b be the permutation of Ω_b such that for $1 \leq j < j' \leq d_b$ the following inequality holds $\beta_{(b,j)} \leq \beta_{(b,j')}$, where $\beta_{(b,j)} := \beta_{\sigma_b(j)}$ and where $d_b = \text{card}(\Omega_b)$ is the number of variables assigned to block b . The integral defined by (6) has the following closed form

$$p(\mathbf{x}_{\{b\}} | \boldsymbol{\theta}_b) = \sum_{j=0}^{d_b} (\beta_{(b,j+1)} - \beta_{(b,j)}) f_b(j; \boldsymbol{\theta}_b), \quad (7)$$

where we define $\beta_{(b,0)} = 0$ and $\beta_{(b,d_b+1)} = 1$. Finally function $f_b(\cdot)$ is defined by

$$f_b(j_0; \boldsymbol{\theta}_b) = \prod_{j=1}^{j_0} \nu_{(b,j)}^{x_{(b,j)}} (1 - \nu_{(b,j)})^{1-x_{(b,j)}} \prod_{j=j_0+1}^{d_b} \lambda_{(b,j)}^{x_{(b,j)}} (1 - \lambda_{(b,j)})^{1-x_{(b,j)}}, \quad (8)$$

where $x_{(b,j)} := x_{\sigma_b(j)}$ denotes the j -th variable (according to permutation σ_b) assigned to block b , $\lambda_{(b,j)} := \lambda_{\sigma_b(j)}$, $\nu_{(b,j)} := \nu_{\sigma_b(j)}$ and where $\prod_{j=j_0+1}^{j_0}$ is one.

The strength of the proposed model is its easy interpretation. The parameter interpretation is allowed by the property of identifiability now presented.

Proposition 2.2 (Model identifiability) The distribution defined by (7) is identifiable under the following constraints:

- $\delta_{(b,1)} = 1$ if $d_b > 2$, $\delta_{(b,1)} = 1$;
- $\varepsilon_{(b,1)} = \varepsilon_{(b,2)}$ if $d_b = 2$;
- $\delta_{(b,1)} = 1$ and $\varepsilon_{(b,1)} = 0$ if $d_b = 1$;

where $\delta_{(b,j)} := \delta_{\sigma_b(j)}$, $\varepsilon_{(b,j)} := \varepsilon_{\sigma_b(j)}$.

The proposed model allows a wide range of dependencies. The following proposition is related the model parameters and Cramer's V. Thus, we can see that the full dependency (respectively anti-dependency) can be modelled by putting $\varepsilon_j = \varepsilon_{j'}$, $\alpha_j = \alpha_{j'}$ and $\delta_j = \delta_{j'}$ (respectively $\varepsilon_j = \varepsilon_{j'}$, $\alpha_j = 1 - \alpha_{j'}$ and $\delta_j = 1 - \delta_{j'}$).

Proposition 2.3 (Dependency measures) The dependency between two binary variables is often measured with Cramer's V. For the distribution defined by (7), Cramer's V between X_j and $X_{j'}$ is zero. Moreover, for j and j' and $\beta_{(b,j)} \leq \beta_{(b,j')}$

$$V(X_j, X_{j'}) = \varepsilon_{(b,j)} \varepsilon_{(b,j')} \sqrt{\frac{\beta_{(b,j)}(1 - \beta_{(b,j')})}{\beta_{(b,j')}(1 - \beta_{(b,j)})}}. \quad (9)$$

3 Parameter inference

3.1 Inference function for Margins

We observed a sample $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ assumed to be composed of n independent realizations of the proposed model. The likelihood related to model $\boldsymbol{\omega}$ is defined by

$$p(\mathbf{x} | \boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{i=1}^n \prod_{b=1}^B p(\mathbf{x}_{i\{b\}} | \boldsymbol{\theta}_b). \quad (10)$$

The *log-likelihood* function is defined by

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\varepsilon}; \mathbf{x}, \boldsymbol{\omega}) = \sum_{i=1}^n \sum_{b=1}^B \ln p(\mathbf{x}_{i\{b\}} | \boldsymbol{\theta}_b), \quad (11)$$

where $\boldsymbol{\alpha} = (\alpha_j; j = 1, \dots, d)$, $\boldsymbol{\delta} = (\delta_j; j = 1, \dots, d)$ and $\boldsymbol{\varepsilon} = (\varepsilon_j; j = 1, \dots, d)$. The proposed distribution is a multivariate copula-based model since each multivariate parametric distribution can be defined as a copula. When the model at hand is a copula with discrete margins, the maximization of the likelihood is quite difficult. Therefore, we use the Inference Function for Margins (IFM) procedure (Joe, 1997). This estimation procedure is based on two optimization steps. The first step maximizes the likelihood of univariate margins. The second step maximizes the dependency parameters with the univariate parameters hold fixed from the first step. Joe (2005) shows the asymptotical efficiency of such a procedure. Thus, the parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\varepsilon}})$ are estimated by the two following steps:

Margin step: for $j \in \{1, \dots, d\}$

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n x_{ij},$$

Dependency step:

$$(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\varepsilon}}) = \arg \max_{(\boldsymbol{\delta}, \boldsymbol{\varepsilon})} L(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}, \boldsymbol{\varepsilon}; \mathbf{x}, \boldsymbol{\omega}).$$

The margin step is easily performed, but the search of $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\varepsilon}})$ at the dependency step requires solving equations having no analytical solution (except when $d_b = 2$). This step is also achieved by using the latent structure of the data when $d_b > 2$ (details are given in Section 3.2). When $d_b = 2$, for $j \in \Omega_b$:

$$\hat{\delta}_{(b,2)} = \begin{cases} 1 & \text{if } n_{11} \geq \hat{\alpha}_{(b,1)} \hat{\alpha}_{(b,2)} \\ 0 & \text{if } n_{11} < \hat{\alpha}_{(b,1)} \hat{\alpha}_{(b,2)} \end{cases} \quad \text{and } \hat{\varepsilon}_{(b,1)} = \hat{\varepsilon}_{(b,2)} = \sqrt{\frac{|n_{11} - \hat{\alpha}_{(b,1)} \hat{\alpha}_{(b,2)}|}{\hat{\beta}_{(b,1)}(1 - \hat{\beta}_{(b,2)})}}, \quad (12)$$

where $n_{11} = \frac{1}{n} \sum_{i=1}^n x_{ij_1} x_{ij_2}$ with $j_1 \in \Omega_b$, $j_2 \in \Omega_b$ and $j_1 \neq j_2$.

3.2 An EM algorithm for the dependency step

Since the blocks of the one-factor distributions imply latent variables, it is natural to perform the dependency steps of the IFM procedure with an Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) when $d_b > 2$. The *complete-data likelihood* is defined by

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{u}, \boldsymbol{\omega}) = \sum_{j=1}^n L(\boldsymbol{\theta}_j; \mathbf{x}, \mathbf{u}, \boldsymbol{\omega}) \quad (13)$$

where

$$\begin{aligned} L(\alpha_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{u}, \boldsymbol{\omega}) = & \sum_{i=1}^n z_{ij} (x_{ij} \ln \lambda_j + (1 - x_{ij}) \ln(1 - \lambda_j)) \\ & + (1 - z_{ij}) (x_{ij} \ln \nu_j + (1 - x_{ij}) \ln(1 - \nu_j)), \end{aligned} \quad (14)$$

where $z_{ij} = 1$ if $0 \leq u_{i\omega_j} < \beta_j$ and $z_{ij} = 0$ if $\beta_j \leq u_{i\omega_j} \leq 1$. The EM algorithm is an iterative algorithm which alternates between two steps: the computation of conditional expectation of the complete-data log-likelihood (E step) and its maximization (M step) on $(\boldsymbol{\delta}, \boldsymbol{\varepsilon})$. Note that the estimate $\hat{\boldsymbol{\alpha}}$ is not modified by the algorithm. Its iteration $[r]$ is written as:

E step: Computation of the complete-data log-likelihood, for $j \in \{1, \dots, d\}$

$$t_{ij}(\boldsymbol{\theta}^{[r]}) = \mathbb{E}[Z_{ij} | \mathbf{x}_i, \boldsymbol{\omega}, \boldsymbol{\theta}_j^{[r]}] = \frac{\lambda_j^{[r]} \beta_j^{[r]}}{\lambda_j^{[r]} \beta_j^{[r]} + \nu_j^{[r]} (1 - \beta_j^{[r]})}. \quad (15)$$

M step: Maximization over $(\delta_j, \varepsilon_j)$, for $j \in \{1, \dots, d\}$

$$\delta_j^{[r+1]} = \mathbb{1}_{\{\max_{\varepsilon_j} L(\hat{\alpha}_j, \delta_j=1, \varepsilon_j; \mathbf{x}, \mathbf{t}^{[r]}, \boldsymbol{\omega}) > \max_{\varepsilon_j} L(\hat{\alpha}_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{t}^{[r]}, \boldsymbol{\omega})\}}, \quad (16)$$

$$\varepsilon_j^{[r+1]} = \arg \max_{\varepsilon_j} L(\hat{\alpha}_j, \delta_j^{[r+1]}, \varepsilon_j; \mathbf{x}, \mathbf{t}^{[r]}, \boldsymbol{\omega}), \quad (17)$$

where $\boldsymbol{\theta}_j^{[r]} = (\hat{\alpha}_j, \delta_j^{[r]}, \varepsilon_j^{[r]})$, $\lambda_j^{[r]} = (1 - \varepsilon_j^{[r]})\hat{\alpha}_j + \varepsilon_j^{[r]\delta_j^{[r]}}$, $\nu_j^{[r]} = (1 - \varepsilon_j^{[r]})\hat{\alpha}_j + \varepsilon_j^{[r]1-\delta_j^{[r]}}$. Thus, the M step involves the search of the maximum over $\varepsilon_j \in]0, 1[$ of $L(\hat{\alpha}_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{t}^{[r]}, \boldsymbol{\omega})$. This maximization is easily performed since it only leads to solve a quadratic equation as shown by Appendix B.

4 Model selection

4.1 Information criterion

Model selection is obviously necessary when we are faced with model-based statistical inference. When the model pmf is given by (2), selecting a model means identifying the repartition of the variables into independent blocks. The challenge also consists of finding the best model according to the data among a set of competing models. In a Bayesian framework, the best model is defined by the model having the highest posterior probability. By assuming that uniformity holds for the prior distribution of $\boldsymbol{\omega}$, the best model also maximizes the integrated likelihood $p(\mathbf{x}|\boldsymbol{\omega})$ where

$$p(\boldsymbol{\omega}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\omega}) \text{ with } p(\mathbf{x}|\boldsymbol{\omega}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\omega}) d\boldsymbol{\theta}, \quad (18)$$

and $p(\boldsymbol{\theta}|\boldsymbol{\omega})$ corresponds to the prior distribution of the parameters of model $\boldsymbol{\omega}$. However, this integral has not a closed form. In this case, the BIC (Schwarz et al., 1978) is used for approaching the logarithm of the integrated likelihood by using a Laplace approximation. It is defined by

$$\text{BIC}(\boldsymbol{\omega}) = L(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}; \mathbf{x}, \boldsymbol{\omega}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln(n), \quad (19)$$

where $\nu_{\boldsymbol{\omega}}$ corresponds to the number of continuous parameters involved in model $\boldsymbol{\omega}$ and where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ is the MLE of model $\boldsymbol{\omega}$. As shown by Gao and Song (2010), the MLE can be replaced in (19) by the estimate provided by the IFM procedure. Thus, we want to obtain model $\boldsymbol{\omega}^*$ which maximizes the BIC criterion among all the competing models.

The number of competing models is too huge for applying an exhaustive approach. Therefore, Section 4.2 presents a deterministic procedure for model selection. This procedure applies a filter among the competing models and only selects d models. Then, the BIC criterion is computed for each of the selected models. We show that this procedure returns the correct model ω^* asymptotically with probability one. Moreover, Section 4.3 presents a stochastic algorithm which finds ω^* . Section 5 shows that both procedures have the same behaviour for detecting the true model, but that the deterministic procedure is drastically faster than the stochastic procedure. Both procedures are implemented in the R package MvBinary, but we advise to only use the deterministic procedure for computing reasons.

4.2 Deterministic approach for model selection

This deterministic procedure has two steps. First, the *reduction step* reduces the number of competing models to only d competing models. Second, the *comparison step* computes the BIC criterion for each of the d resulting models and the model maximizing the BIC criterion is returned.

The reduction step decreases the number of competing models by using the empirical dependencies between the variables. Indeed, it performs the Hierarchical Ascendant Classification (HAC) of the variables based on the empirical Cramer's V. This procedure proposes d partitions corresponding to the d competing models on which the BIC criterion will be computed. Each model proposed by the HAC is relevant since it models the strongest empirical dependencies. Moreover, the HAC provides embedded partitions of variables and then reduces the calls to the EM algorithm.

The deterministic procedure based on HAC performs the model selection with the two following steps:

Reduction step performs the HAC based on the empiric Cramer's V to defined the d partitions of the variables.

Comparison step computes $\text{BIC}(\omega^{[k]})$ for $k = 1, \dots, d$, where $\omega^{[k]}$ is such that each block b is composed by the variables affiliated to class b by the partition into k classes of the HAC.

The procedure returns $\arg \max_{k=1, \dots, d} \text{BIC}(\omega^{[k]})$.

Proposition 4.1 (Consistency of the HAC-based procedure) *The HAC-based procedure is asymptotically consistent (i.e. it returns the true model with probability one when n grows to infinity).*

Proof is given in Appendix B.

4.3 Stochastic approach for model selection

Model ω^* can be determined through a Metropolis-Hastings algorithm (Robert and Casella, 2004). This algorithm performs a random walk over the competing models and its unique invariant distribution is proportional to $\exp(\text{BIC}(\omega))$. Therefore, ω^* is the mode of its stationary distribution. It is also sampled with probability one by the algorithm when the number of iterations R grows to infinity.

At iteration $[r]$, the algorithm samples a model candidate $\tilde{\omega}$ from the distribution $q(\cdot|\omega^{[r]})$ where $\omega^{[r]}$ corresponds to the current model. More precisely, candidate $\tilde{\omega}$ is equal to the current model $\omega^{[r]}$ except for variable $j^{[r]}$ randomly sampled which is affiliated into block $b^{[r]}$ randomly sampled in $\{1, \dots, \max(\omega^{[r]}) + 1\}$. This candidate is accepted with a probability equal to

$$\rho^{[r]} = \frac{\exp(\text{BIC}(\tilde{\omega}))q(\omega^{[r]}|\tilde{\omega})}{\exp(\text{BIC}(\omega^{[r]}))q(\tilde{\omega}|\omega^{[r]})}. \quad (20)$$

This algorithm performs R iterations and returns the model maximizing the BIC criterion. In practice, there may be almost absorbing states, so different initialisations of this algorithm ensure to visit ω^* . Thus, starting from $\omega^{[0]}$, uniformly sampled among the competing models, the algorithm performs R iterations and returns $\arg \max_{r=1, \dots, R} \text{BIC}(\omega^{[r]})$. Its iteration $[r]$ performs the two following steps:

Candidate step: $\tilde{\omega}$ is sampled from $q(\cdot|\omega^{[r]})$.

Acceptance/reject step: defined $\omega^{[r]}$ with

$$\omega^{[r]} = \begin{cases} \tilde{\omega} & \text{with probability } \rho^{[r]} \\ \omega^{[r-1]} & \text{otherwise} \end{cases}.$$

5 Numerical experiments

5.1 Suitability of the HAC-based procedure

This simulation shows the relevance of competing models provided by the reduction step of the HAC-based procedure. Data are simulated from the proposed model with the following parameters

$$d = 10, \delta_j = 1, \alpha_j = 0.4 \text{ and } \omega_j = \begin{cases} 1 & \text{if } 1 \leq j \leq 5 \\ 2 & \text{if } 6 \leq j \leq 10 \end{cases}. \quad (21)$$

For different sizes of sample n and strengths of dependencies ε_j , we check if the true model belongs to the list of models returned by reduction step of the HAC-based procedure. Table 1 shows the results obtained on 50 samples for different values of (n, ε_j) .

$n \varepsilon_j$	0.2	0.3	0.4	0.5	0.6
50	0	1	4	28	37
100	1	1	20	41	49
200	0	9	40	48	20
400	2	29	47	50	50
800	3	44	50	50	50
1600	21	49	50	50	50
3200	40	50	50	50	50

Table 1: Number of times where the true model belongs to the list of models returned by the reduction step of the HAC-based procedure on 50 samples.

Thus, whatever the strength of dependencies, the procedure asymptotically proposes the true model. Obviously, for a fixed sample size, results are better when the dependencies are strong since the number of times where the true model belongs to the list of models is increasing with the dependency strength.

5.2 Comparison of model selection procedures

Both procedures of model selection are compared on data sampled from the proposed model with the parameters defined in (21). To compare the quality of the estimates returned by both procedures, we use the Kullback-Leibler divergence. As shown by Table 2, both procedures are consistent since the Kullback-Leibler divergence asymptotically vanishes. Moreover, the estimates have the same quality (equal value of the Kullback-Leibler divergence). However, the HAC-based procedure is considerably faster than the Metropolis-Hastings procedure as shown by Table 3. So, we recommend to use the HAC-based procedure to perform the model selection in high dimension.

$n \varepsilon_j$	0.2		0.3		0.4		0.5	
	HAC	MH	HAC	MH	HAC	MH	HAC	MH
50	0.15	0.16	0.24	0.27	0.36	0.36	0.39	0.46
100	0.08	0.09	0.13	0.14	0.22	0.21	0.12	0.17
200	0.04	0.04	0.10	0.10	0.08	0.11	0.05	0.06
400	0.03	0.03	0.06	0.06	0.03	0.03	0.03	0.03

Table 2: Kullback-Leibler divergence obtained with the estimates provided by both procedure of model selection.

$n \varepsilon_j$	0.2		0.3		0.4		0.5	
	HAC	MH	HAC	MH	HAC	MH	HAC	MH
50	11	217	10	250	9	278	8	381
100	12	241	11	250	10	354	8	633
200	14	276	13	308	11	662	9	912
400	16	296	15	509	12	1218	9	933

Table 3: Computing time in seconds required by the two procedures of model selection.

5.3 Model selection for high-dimensional data

This section shows the behaviour of the HAC-based procedure in high dimension. Data are generated from a model with blocks of five dependent variables ($d_b = 1$), with equal marginal probabilities ($\alpha_j = 0.4$) and equal dependency strength ($\varepsilon_j = 0.4$ and $\delta_j = 1$). For different sizes of sample and numbers of variables, 50 samples are generated.

Table 4 shows the relevance of the deterministic procedure by using the Adjusted Rand Index (ARI) to compare the true partition of the variables into blocks and its

estimated. Indeed, whatever the number of variables, the procedure provides the true model with probability one when n grows to infinity. However, for small samples the procedure can provide a model slightly different to the true model.

$n d$	10	20	50	100	200
50	0.11 (0.18)	0.07 (0.03)	0.10 (0.06)	0.10 (0.04)	0.06 (0.02)
100	0.35 (0.33)	0.35 (0.24)	0.24 (0.14)	0.22 (0.07)	0.15 (0.03)
200	0.85 (0.27)	0.78 (0.20)	0.67 (0.11)	0.56 (0.07)	0.43 (0.05)
400	0.97 (0.09)	0.95 (0.07)	0.95 (0.05)	0.91 (0.05)	0.86 (0.05)
800	1.00 (0.00)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)

Table 4: Mean (in bold) and standard deviation (in parenthesis) of the ARI between ω^0 and ω^* .

6 Application to plant distribution in the USA

Dataset

Data has been extracted from the USA plants database, July 29, 2015. It describes 35583 plants by indicating if they occur in 69 states (USA, Canada, Puerto Rico, Virgin Islands, Greenland and St Pierre and Miquelon). By modelling the data distribution, the flora variety of each states could be characterized. Moreover, one can expect bring out geographic dependencies between the variables. The data are available in the R package MvBinary which implements the proposed method.

Experiment conditions

The model selection is achieved by the deterministic algorithm (see Section 4.2) where the Ward criterion is used for the HAC. The EM algorithm is randomly initialized 40 times and it is stopped when two successive iterations increase the log-likelihood less than 0.01.

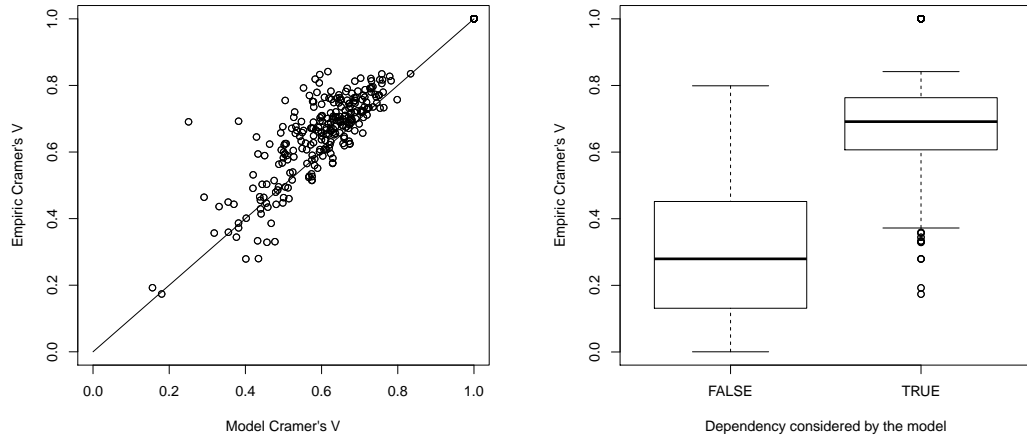
Model coherence

Figure 1 shows the relevance of the dependencies detected by the estimated model. Indeed, Figure 1a shows the correspondence between Cramer’s V computed with the model parameter and the empirical Cramer’s V, for each pair of variables claimed to be dependent by the estimated model. Moreover, Figure 1b shows that the estimated model well represents the main dependencies.

The estimated model is composed of 10 blocks of dependent variables. Figure 2 shows that this block repartition has a geographic meaning.

Model interpretation

Parameters permit an easy interpretation of the whole distribution. The mean per block of the values of $\hat{\alpha}_j$ and $\hat{\varepsilon}_j$ are summarized by Figure 3. Note that the dependencies



(a) Empiric Cramer's V and Cramer's V estimated by the model. (b) Boxplot of the Empiric Cramer's V for the modelled and not modelled dependencies

Figure 1: Visualisation of the dependencies taken into account by the model.

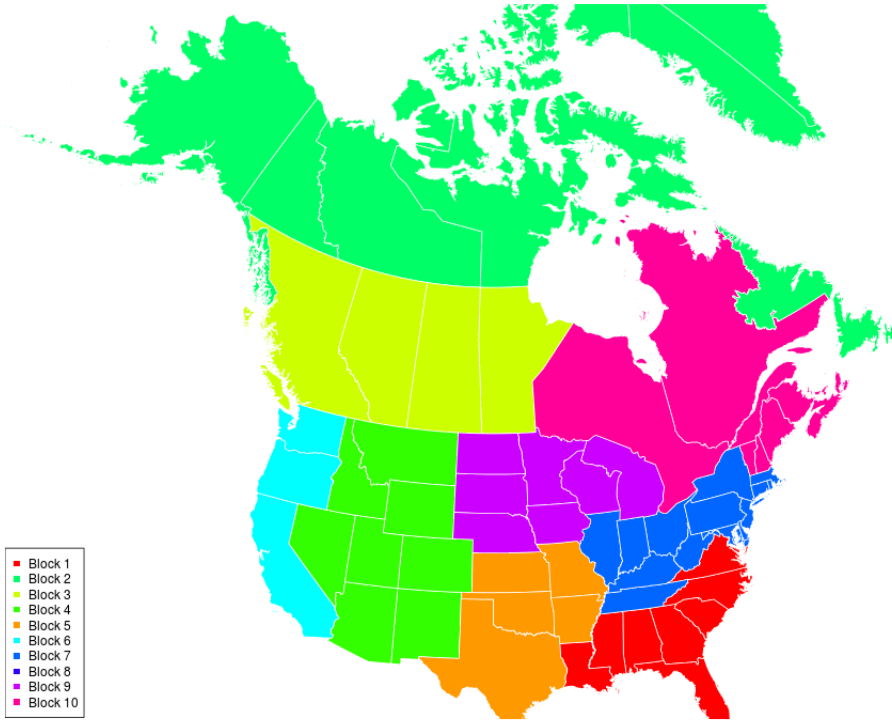


Figure 2: Geographic coherence of the blocks of states (color indicates the block assignment)

detected by the model are all positive since for $j = 1, \dots, d$, $\hat{\delta}_j = 1$.

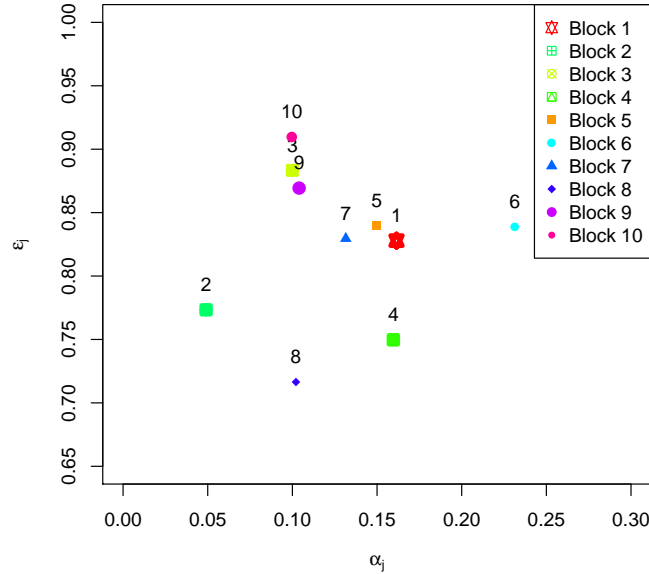


Figure 3: Summary of the parameters by blocks

Each block is composed of highly dependent variables (high values of parameters $\hat{\epsilon}_j$ and $\hat{\delta}_j = 1$). Therefore, the knowledge of one variables of a block provides strong information about the other variables affiliated into this block. For instance, the most dependent block is Block 10 (composed by Prince Edward Island, Nova Scotia, New Brunswick, New Hampshire, Vermont, Maine, Québec and Ontario). Thus, a plant occurs in Ontario with probability $\hat{\alpha}_{Ontario} = 0.14$ while it occurs with a probability 0.83 if this plant occurs in Québec. The least dependent block is composed of tropical states (Virgin Islands, Puerto Rico and Hawaii). These weaker dependencies can be explained by large geographic distance. Finally, parameters α_j allow to described the region by their amount of plants. Cold regions (Blocks 2, 3 and 10) obtains small values of $\hat{\alpha}_j$ while the "sun-belt" obtains large values of this parameter.

7 Conclusion

In this paper, we have introduced a new family of distributions for large binary datasets. This family implies that the variables are grouped into independent blocks and that each block follows a specific one factor distribution. This new family has many good properties. Indeed, it verifies the five features required by Nikoloulopoulos (2013) for a "good" distribution. Moreover, it permits an easy interpretation of the whole distribution. The variable repartition puts the light on the main dependencies. Moreover, each variable is summarized by its marginal probability (parameter α_j) and by its strength (parameter ϵ_j) and its kind (parameter δ_j) of dependency with the other block variables. Finally,

this model is suitable for modelling large binary data since its number of parameters is linear in d .

We have proposed to circumvent the combinatorial problem of model selection with a deterministic procedure which reduces the number of competing models by using the empirical dependencies. Although this procedure does not ensure the maximization of the BIC, its consistency has been demonstrated. Numerical experiments have shown that this approach provides estimates having the same quality as a stochastic (and optimal) procedure, but it strongly reduces the computing time. The R package MvBinary implements both procedures of inference and contains the data set used in the application.

Many extension of this work can be envisaged. Indeed, parsimony extensions could be introduced by imposing equality constraints between the block parameters (*e.g.* $\forall j \in \Omega_b, \varepsilon_j = c_b$ where $c_b \in]0, 1]$). Moreover, more complex dependencies could be modelled by considering more than one factor and by keeping the same kind of parametrization. However, the parameter estimation and the likelihood computation would be more complex. Indeed, the pmf of block b would be defined as a sum of $(d_b + 1)^2$ terms, while it is currently a sum of $d_b + 1$ terms.

Finally, this model could be an answer to difficult task of the binary data clustering with intra-component dependencies. Indeed, the clustering aim could be achieved by considering a finite mixture of the proposed distribution. However, the challenge of model selection would be a complex issue. Moreover, the model identifiability should be carefully studied.

A Proofs of the model properties

Proof of Proposition 2.1 It suffices to remark that (6) can be decomposed into $d_b + 1$ integrals whose bounds are given by the coefficients $\beta_{(b,j)}$. By using the conditional independence between variables given in (3) and the conditional distribution of x_j given by (5), function $p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b)$ is a piecewise constant function of u_b . Thus, for $u_b \in [\beta_{(b,j)}, \beta_{(b,j+1)}[$, $p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b)$ is constant and equal to $f_b(j)$ defined by (8). Then,

$$\begin{aligned} p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) &= \int_0^{\beta_{(b,1)}} p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) du + \sum_{j=1}^{d_b-1} \int_{\beta_{(b,j)}}^{\beta_{(b,j+1)}} p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) du + \int_{\beta_{(b,d_b)}}^1 p(\mathbf{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) du \\ &= \beta_{(b,1)} f_b(0; \boldsymbol{\theta}_b) + \sum_{j=1}^{d_b-1} (\beta_{(b,j+1)} - \beta_{(b,j)}) f_b(j; \boldsymbol{\theta}_b) + (1 - \beta_{(b,d_b)}) f_b(d_b; \boldsymbol{\theta}_b). \end{aligned}$$

Proof of Proposition 2.2 We define that the distribution is identifiable if for two vectors of parameters $\boldsymbol{\theta}_b = (\alpha_j, \varepsilon_j, \delta_j; j \in \Omega_b)$ and $\boldsymbol{\theta}'_b = (\alpha'_j, \varepsilon'_j, \delta'_j; j \in \Omega_b)$ such that

$$\forall \mathbf{x}_{\{b\}}, p(\mathbf{x}_{\{b\}}|\boldsymbol{\theta}_b) = p(\mathbf{x}_{\{b\}}|\boldsymbol{\theta}'_b) \text{ then } \boldsymbol{\theta}_b = \boldsymbol{\theta}'_b. \quad (22)$$

Without loss of generality, we assume that $\alpha_j \leq \alpha_{j+1}$. The equality $\alpha_j = \alpha_{j'}$ is directly obtained since $\forall j \in \Omega_b, \alpha_j = p(x_j = 1|\boldsymbol{\theta}_b) = p(x_j = 1|\boldsymbol{\theta}'_b) = \alpha'_j$. The proof distinguishes three cases: one variable in the block (*i.e.* $d_b = 1$) with the constraints $\delta_{(b,1)} = 1$ and $\varepsilon_{(b,1)} = 0$; two variables in the block (*i.e.* $d_b = 2$) with the constraints $\delta_{(b,1)} = 1$ and

$\varepsilon_{(b,1)} = \varepsilon_{(b,2)}$; more than two variables in the block (*i.e.* $d_b > 2$) with the constraint $\delta_{(b,1)} = 1$. Proofs use the following probability: $\forall (j_1, j_2) \in \Omega_b$,

$$p(x_{j_1} = 1, x_{j_2} = 1 | \theta_b) = \begin{cases} \alpha_{j_1} \alpha_{j_2} + \varepsilon_{j_1} \varepsilon_{j_2} \alpha_{j_1} (1 - \alpha_{j_2}) & \text{if } \delta_{j_2} = 1 \\ \alpha_{j_1} \alpha_{j_2} - \varepsilon_{j_1} \varepsilon_{j_2} \alpha_{j_1} \alpha_{j_2} & \text{if } \delta_{j_2} = 0 \text{ and } \alpha_{j_1} + \alpha_{j_2} < 1 \\ \alpha_{j_1} \alpha_{j_2} - \varepsilon_{j_1} \varepsilon_{j_2} (1 - \alpha_{j_1})(1 - \alpha_{j_2}) & \text{if } \delta_{j_2} = 0 \text{ and } \alpha_{j_1} + \alpha_{j_2} \geq 1 \end{cases} \quad (23)$$

If $\delta_{(b,j)} \neq \delta'_{(b,j)}$ then without loss of generality we assume that $\delta_{(b,j)} = 1$ and $\delta'_{(b,j)} = 0$. From (23), $p(x_1 = 1, x_j = 1 | \theta_b) > \alpha_{(b,1)} \alpha_{(b,j)} = \alpha'_{(b,1)} \alpha'_{(b,j)} > p(x_1 = 1, x_j = 1 | \theta'_b)$ but this is in contradiction to (22), so $\forall j \in \Omega_b$, $\delta'_{(b,j)} = \delta_{(b,j)}$. Therefore, we have to prove the equality $\varepsilon_{(b,j)} = \varepsilon'_{(b,j)}$.

Case 1 ($d_b = 1$ with $\delta_{(b,1)} = 1$ and $\varepsilon_{(b,1)} = 0$). Then parametrization assumes that only parameter α is free. Equality $\alpha_j = \alpha'_j$ implies $\theta_b = \theta'_b$.

Case 2 ($d_b = 2$ with $\delta_{(b,1)} = 1$ and $\varepsilon_{(b,1)} = \varepsilon_{(b,2)}$). By using constraints $\varepsilon_{(b,1)} = \varepsilon_{(b,2)}$ and $\varepsilon'_{(b,1)} = \varepsilon'_{(b,2)}$ and by using (23), then $\varepsilon_{(b,1)}^2 = \varepsilon_{(b,1)}'^2$. Thus, $\theta_b = \theta'_b$.

Case 3 ($d_b > 2$ with $\delta_{(b,1)} = 1$). (23) is verified by θ and θ' . Moreover, we know that $\alpha_j = \alpha'_j$ and $\delta_j = \delta'_j$, for $j = 1, \dots, d$. So, the following system S arises from (23) for $(j_1, j_2) = \{(1, 2), (1, 3), \dots, (1, d_b), (2, 3)\}$

$$(S) = \begin{cases} \varepsilon_{(b,1)} \varepsilon_{(b,2)} &= \varepsilon'_{(b,1)} \varepsilon'_{(b,2)} \\ \varepsilon_{(b,1)} \varepsilon_{(b,3)} &= \varepsilon'_{(b,1)} \varepsilon'_{(b,3)} \\ \vdots &\vdots \\ \varepsilon_{(b,1)} \varepsilon_{(b,d_b)} &= \varepsilon'_{(b,1)} \varepsilon'_{(b,d_b)} \\ \varepsilon_{(b,2)} \varepsilon_{(b,3)} &= \varepsilon'_{(b,2)} \varepsilon'_{(b,3)} \end{cases} \quad (24)$$

If $\varepsilon'_{(b,1)} \neq \varepsilon_{(b,1)}$ then $\exists t \neq 1$ such that $\varepsilon'_{(b,1)} = t \varepsilon_{(b,1)}$. Then, the first d_b lines of (S) imply that $\forall j = 2, \dots, d_b$, $\varepsilon_{(b,j)} = t \varepsilon'_{(b,j)}$. Since the last line of (S) implies that $\varepsilon_{(b,2)} \varepsilon_{(b,3)} = \varepsilon_{(b,2)} \varepsilon_{(b,3)} / t^2$, positivity of $\varepsilon_{(b,j)}$ permits to conclude that $\varepsilon'_{(b,1)} = \varepsilon_{(b,1)}$, so $\forall j = 2, \dots, d_b$, $\varepsilon_{(b,j)} = \varepsilon'_{(b,j)}$. Thus, $\theta_b = \theta'_b$.

Proof of Proposition 2.3 We denote $p_{hh'} = P(X_{\sigma_b(j)} = h, X_{\sigma_b(j')} | \omega, \theta)$ with $j < j'$ and $h \in \{0, 1\}$ and $h' \in \{0, 1\}$. Then

$$\begin{aligned} p_{11} &= \alpha_{(b,j)} \alpha_{(b,j')} + r \\ p_{01} &= (1 - \alpha_{(b,j)}) \alpha_{(b,j')} - r \\ p_{10} &= \alpha_{(b,j)} (1 - \alpha_{(b,j')}) - r \\ p_{00} &= (1 - \alpha_{(b,j)}) (1 - \alpha_{(b,j')}) + r \end{aligned}$$

where $r = \varepsilon_{(b,j)} \varepsilon_{(b,j')} \beta_{(b,j)} (1 - \beta_{(b,j')})$. Thus, (9) is obtained by applying the definition of Cramer's V.

B Details about the M-step of the EM algorithm

By using the definition of $\hat{\alpha}_j$, $\hat{\alpha}_j = n_{10} + n_{01}$ where $n_{10} = \frac{1}{n} \sum_{i=1}^n x_{ij} (1 - t_{ij})^{\delta_j} (t_{ij})^{1-\delta_j}$ and $n_{11} = \frac{1}{n} \sum_{i=1}^n x_{ij} (t_{ij})^{\delta_j} (1 - t_{ij})^{\delta_j}$. Moreover, the expectation of the complete-data

likelihood related to variable j is written as

$$L(\hat{\alpha}_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{t}, \boldsymbol{\omega}) = n_{10} \ln((1 - \varepsilon_j)(n_{11} + n_{10})) + n_{11} \ln((1 - \varepsilon_j)(n_{11} + n_{10}) + \varepsilon_j) \quad (25) \\ + n_{00} \ln(1 - (1 - \varepsilon_j)(n_{11} + n_{10})) + n_{01} \ln(1 - (1 - \varepsilon_j)(n_{11} + n_{10}) - \varepsilon_j),$$

where $n_{00} = \frac{1}{n} \sum_{i=1}^n (1 - x_{ij})(1 - t_{ij})^{\delta_j} (t_{ij})^{1-\delta_j}$ and $n_{01} = \frac{1}{n} \sum_{i=1}^n (1 - x_{ij})(t_{ij})^{\delta_j} (1 - t_{ij})^{\delta_j}$. For a fixed value of δ_j , the argmax over ε_j of $L(\hat{\alpha}_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{t}, \boldsymbol{\omega})$ is denoted by $\varepsilon_{j|\delta_j}$. The estimation of $\varepsilon_{j|\delta_j}$ is obtained by setting to zero the derivative of $L(\hat{\alpha}_j, \delta_j, \varepsilon_j; \mathbf{x}, \mathbf{t}, \boldsymbol{\omega})$ over ε_j . So, remarking that $n_{01} = 1 - n_{11} - n_{10} - n_{00}$,

$$\frac{n_{11} + n_{00} - 1}{1 - \varepsilon_{j|\delta_j}} + \frac{n_{11}(1 - n_{11} - n_{10})}{n_{11} + n_{10} + \varepsilon_{j|\delta_j}(1 - n_{11} - n_{10})} + \frac{n_{01}(n_{11} + n_{10})}{(n_{11} + n_{10})\varepsilon_{j|\delta_j} + (1 - n_{11} - n_{10})} = 0. \quad (26)$$

This equation is equivalent to the following quadratic equation

$$\varepsilon^2 A + \varepsilon B + C = 0, \quad (27)$$

where $A = -(n_{11} + n_{10})(1 - n_{11} - n_{10})$, $B = n_{11}(n_{11} + n_{10}) + n_{00}(1 - n_{11} - n_{10}) - (n_{11} + n_{10})^2 - (1 - n_{11} - n_{10})^2$ and where $C = n_{11}(1 - n_{11} - n_{10}) + n_{00}(n_{11} + n_{10}) + A$. Let s_1 and s_2 be the two solutions of (27):

$$s_1 = \frac{-B - \sqrt{\Delta}}{2A} \text{ and } s_2 = \frac{-B + \sqrt{\Delta}}{2A}, \quad (28)$$

where $\Delta = B^2 - 4AC$. By noting that $\varepsilon_j \in]0, 1[$, and that $s_1 = \frac{(n_{11} + n_{10})n_{10} + (1 - n_{11} - n_{10})n_{01}}{-2(n_{11} + n_{10})(1 - n_{11} - n_{10})} < 0$, we conclude that $\varepsilon_{j|\delta_j} = \max(0, s_2)$.

Consistency of the HAC-based procedure

The proof of Proposition 4.1 is done in three steps. First, we show that the HAC-based procedure applied to the theoretical Cramer's matrix is consistent. Second, we show that this result holds in a neighbourhood of the theoretical Cramer's matrix. Third, we conclude by using the convergence in probability of the empiric Cramer's matrix to the theoretical one.

Let $M^0 \in [0, 1]^{d \times d}$ be the dissimilarity matrix based on Cramer's V computed with the true distribution defined by model $\boldsymbol{\omega}^0$ and its parameters $\boldsymbol{\theta}^0$. So, for $1 \leq j, j' \leq d$

$$M^0(j, j') = 1 - V^0(X_j, X_{j'}) \quad (29)$$

with $V^0(X_j, X_{j'})$ is the theoretical Cramer's V between X_j and $X_{j'}$ defined by

$$V^0(X_j, X_{j'}) = \sqrt{\sum_{h=0}^1 \sum_{h'=0}^1 \frac{(P(X_j = h, X_{j'} = h'; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0) - P(X_j = h; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0)P(X_{j'} = h'; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0))^2}{P(X_j = h; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0)P(X_{j'} = h'; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0)}}, \quad (30)$$

Since the true model $\boldsymbol{\omega}^0$ involves independence between blocks of variables, for $1 \leq j, j' \leq d$ with $\omega_j^0 \neq \omega_{j'}^0$, $M^0(j, j') = 1$. We denote by μ^0 the greatest value of M^0 when the variables belong to the same block for the true model $\boldsymbol{\omega}^0$

$$\mu^0 = \arg \max_{\{(j, j') : \omega_j^0 = \omega_{j'}^0\}} M^0(j, j'). \quad (31)$$

Note that $\mu^0 < 1$ since the variables affiliated into the same block are dependent. Finally, $\mathbf{\Omega}^{[r]} = (\mathbf{\Omega}_b^{[r]}; b = 1, \dots, d)$ denotes the partition provided by the HAC at its iteration $[r]$, where $\mathbf{\Omega}_b^{[r]}$ is the set of the indices of the variables affiliated to block b at iteration $[r]$. We consider that the HAC is used with a classical dissimilarity measure $D(.,.)$ (min, max, mean or Ward).

Proposition B.1 *If $\exists(j_1, j_2) \in \{1, \dots, d\}^2$ with $\omega_{j_1}^0 = \omega_{j_2}^0$ and with $j_1 \in \mathbf{\Omega}_{b_1}^{[r]}$, $j_2 \in \mathbf{\Omega}_{b_2}^{[r]}$ and $b_1 \neq b_2$, then*

$$\forall b, \forall(j, j') \in \mathbf{\Omega}_b^{[r+1]} : \omega_j^0 = \omega_{j'}^0. \quad (32)$$

Proof At iteration $[0]$, each variable is affiliated into its own block, so $\mathbf{\Omega}_b^{[r]} = \{b\}$ for $b \in \{1, \dots, d\}$. Let $(j_1^{[0]}, j_2^{[0]}) = \arg \min_{(j_1, j_2)} M^0(j_1, j_2)$, then

$$\mathbf{\Omega}_b^{[1]} = \begin{cases} \mathbf{\Omega}_b^{[0]} & \text{if } b \neq j_1^{[0]} \text{ and } b \neq j_2^{[0]} \\ \mathbf{\Omega}_{j_1^{[0]}}^{[0]} \cup \mathbf{\Omega}_{j_2^{[0]}}^{[0]} & \text{if } b = j_1^{[0]} \\ \emptyset & \text{if } b = j_2^{[0]} \end{cases}. \quad (33)$$

The $\mathbf{\Omega}^{[1]}$ verifies (32).

At iteration $[r]$, by definition $\forall b, \forall(j, j') \in \mathbf{\Omega}_b^{[r]} : \omega_j^0 = \omega_{j'}^0$. Let the couple $(b_1^{[r]}, b_2^{[r]}) = \arg \min_{(b_1, b_2) \text{ with } b_1 \neq b_2} D(\mathbf{\Omega}_{b_1}^{[r]}, \mathbf{\Omega}_{b_2}^{[r]})$. There are two cases to be considered, for all $j_1 \in \mathbf{\Omega}_{b_1}^{[r]}$ and $j_2 \in \mathbf{\Omega}_{b_2}^{[r]}$,

- if $\omega_{j_1}^0 \neq \omega_{j_2}^0$ then $D(\mathbf{\Omega}_{b_1}^{[r]}, \mathbf{\Omega}_{b_2}^{[r]}) = 1$.
- if $\omega_{j_1}^0 = \omega_{j_2}^0$ then $D(\mathbf{\Omega}_{b_1}^{[r]}, \mathbf{\Omega}_{b_2}^{[r]}) \leq \mu^0$.

Since $\mu^0 < 1$, (32) is verified.

Corollary B.2 (Consistency with theoretical matrix) *The HAC based on the dissimilarity matrix M provides the true model at its iteration $d - B^0$ where B^0 is the number of blocks defined by ω^0 .*

Proof It is the only partition of $B^{[0]}$ classes which respects Proposition B.1.

Corollary B.3 (Consistency in a neighbourhood of the theoretical matrix) *The HAC based on dissimilarity matrix M belonging to a neighbourhood of M^0 , denoted by $N(M^0)$, provides the true model at its iteration $d - B^0$ where*

$$N(M^0) = \left\{ M \in [0, 1]^{d \times d} \text{ with } |M(j, j') - M^0(j, j')| < \frac{1 - \mu^0}{2} \right\}. \quad (34)$$

Proof Proof is based on the same reasoning as the proof of Proposition B.1, since we have

$$\begin{cases} M(j, j') > \frac{1 + \mu^0}{2} & \text{if } \omega_j^0 \neq \omega_{j'}^0 \\ M(j, j') < \frac{1 + \mu^0}{2} & \text{if } \omega_j^0 = \omega_{j'}^0 \end{cases}.$$

Proof of Proposition 4.1 The Law of Large numbers implies that the observed probability of each couple (j, j') converges in probability to its true value: for any $h \in \{0, 1\}$ and $h' \in \{0, 1\}$

$$\hat{p}_{hh'} \xrightarrow{pr} P(X_j = 1, X_{j'} = 1; \boldsymbol{\omega}^0, \boldsymbol{\theta}^0), \quad (35)$$

where $p_{hh'} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_{ij}=h} \mathbb{1}_{x_{ij'}=h'}$.

The empirical Cramer's V denoted by \hat{V} is a continuous function of $\hat{p}_{hh'}$, since it is defined by

$$\hat{V}(X_j, X_{j'}) = \sqrt{\sum_{h=0}^1 \sum_{h'=0}^1 \frac{(\hat{p}_{hh'} - \hat{p}_{h\bullet} \hat{p}_{\bullet h'})^2}{\hat{p}_{h\bullet} \hat{p}_{\bullet h'}}}, \quad (36)$$

where $\hat{p}_{h\bullet} = \hat{p}_{h0} + \hat{p}_{h1}$ and $\hat{p}_{\bullet h'} = \hat{p}_{0h'} + \hat{p}_{1h'}$. Thus, the Mapping theorem (see for instance Theorem 2.7 page 21 of Billingsley (2013)) implies that \hat{V} converges in probability to V^0 . So,

$$P(M \in N(M^0)) \xrightarrow{n \rightarrow \infty} 1. \quad (37)$$

Thus, by applying Corollary B.3, the probability that $\boldsymbol{\omega}^0$ belongs to the model subset provided by the HAC procedure is equal to one. The consistency of the BIC criterion concludes the proof.

References

- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.
- Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of classification*, 8(2):157–176.
- Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467.
- Collett, D. (2002). *Modelling binary data*. CRC press.
- Cox, D. R. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2):113–120.
- Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data*, volume 32. CRC Press.
- Cox, D. R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika*, 81(2):403–408.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38.
- Gaetan, C., Guyon, X., and Bleakley, K. (2010). *Spatial statistics and modeling*, volume 271. Springer.

- Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.
- Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(02):475–515.
- Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., Fortin, M., et al. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27(3):265–284.
- Glonek, G. F. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the royal statistical society. Series B (Methodological)*, 57(3):533–546.
- Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.
- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.
- Knott, M. and Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Number 7. Edward Arnold.
- Krupskii, P. and Joe, H. (2015). Structured factor copula models: theory, inference and computation. *J. Multivariate Anal.*, 138:53–73.
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50.
- Kurowicka, D. (2011). *Dependence modeling: vine copula handbook*. World Scientific.
- Mazo, G., Girard, S., and Forbes, F. (2015). A flexible and tractable class of one-factor copulas. *Statistics and Computing*, pages 1–15.
- McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, second edition.
- Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203.
- Nelsen, R. B. (2006). *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Nikoloulopoulos, A. K. (2013). *Copula-based models for multivariate discrete response data*. Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics, Springer-Verlag Berlin Heidelberg.
- Nikoloulopoulos, A. K. and Joe, H. (2013). Factor copula models for item response data. *Psychometrika*, 80(1):126–150.

- Nikoloulopoulos, A. K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, 27(30):6393–6406.
- Nikoloulopoulos, A. K. and Karlis, D. (2009). Finite normal mixture copulas for multivariate discrete data modeling. *J. Statist. Plann. Inference*, 139(11):3878–3890.
- Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *J. Amer. Statist. Assoc.*, 107(499):1063–1072.
- Pettitt, A. N., Weir, I. S., and Hart, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. Comput.*, 12(4):353–367.
- Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.
- Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.
- Weir, I. S. and Pettitt, A. N. (2000). Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *J. Roy. Statist. Soc. Ser. C*, 49(4):473–484.